Human Centric Facial Expression Recognition

K. Clawson [1*], L. S. Delicato, [2**] and C. Bowerman, [1***]
[1] Faculty of Computer Science, University of Sunderland, Sunderland, SR1 3SD, UK
[2.] Faculty of Health, Sciences and Wellbeing, University of Sunderland, SR1 3QR, UK
[*] kathy.clawson@sunderland.ac.uk, [**]louise.delicato@sunderland.ac.uk,
***chris.bowerman@sunderland.ac.uk

Facial expression recognition (FER) is an area of active research, both in computer science and in behavioural science. Across these domains there is evidence to suggest that humans and machines find it easier to recognise certain emotions, for example happiness, in comparison to others. Recent behavioural studies have explored human perceptions of emotion further, by evaluating the relative contribution of features in the face when evaluating human sensitivity to emotion. It has been identified that certain facial regions have more salient features for certain expressions of emotion, especially when emotions are subtle in nature. For example, it is easier to detect fearful expressions when the eyes are expressive. Using this observation as a starting point for analysis, we similarly examine the effectiveness with which knowledge of facial feature saliency may be integrated into current approaches to automated FER. Specifically, we compare and evaluate the accuracy of 'full-face' versus upper and lower facial area convolutional neural network (CNN) modelling for emotion recognition in static images, and propose a human centric CNN hierarchy which uses regional image inputs to leverage current understanding of how humans recognise emotions across the face. Evaluations using the CK+ dataset demonstrate that our hierarchy can enhance classification accuracy in comparison to individual CNN architectures, achieving overall true positive classification in 93.3% of cases.

Facial Expression Recognition, Emotion Recognition, Deep Learning, Convolutional Neural Network.

1. INTRODUCTION

Human expression recognition is a challenging topic that plays an increasingly important role across a variety of domains, including psychology, cognitive science, and computer science. The ability to recognize facial expressions automatically offers opportunities for novel applications in human computer interaction, including adaptive user interaction, usability testing, mood tracking, and for the development of systems focusing on enhancing an individual's emotional intelligence.

Traditionally, facial expression recognition (FER) pipelines have incorporated a number of processing stages, such as image or video pre-processing (for example face detection and intensity normalisation), hand crafted feature extraction, feature space representation, and classification. Feature detection for FER has focused on both isolation of facial action units, and on the quantification of more general geometric or appearance based features.

With the recent application of deep learning, FER has made significant progress. Several works focus on the use of Convolutional Neural Networks (CNNs) for collective feature extraction and detection of prototypical emotions (Dachapally, 2017; Xie and Hu, 2017). CNNs constitute an end to end model which take an image input and perform combined feature extraction and classification within a single stage (Liu, Zhang and Pan, 2016). CNNs have been successfully applied, and achieved state of the art performance, for numerous computer vision applications including object detection, (Cireşan, Meier and Schmidhuber, 2012), and face detection and verification (Sun, Wang and Tang, 2014).

Despite achieving state of the art performance, CNNs for FER are subject to a number of constraints. Within the literature, there are differences in evaluation methodologies such as the number of classes considered, and ratio of training to test set samples. Furthermore, it has been reported that not all studies ensure there is no subject overlap between training and test partitions (Lopes *et al.*, 2017). Such factors can lead to misleadingly high cited accuracies, and makes it

Published by BCS Learning and Development Ltd.
Proceedings of British HCI 2018. Belfast, UK
.

difficult to compare competing approaches. Additional constraints include the existence of inter and intra subject variability (instances of expressions can show differences even with the same participant), variations in brightness, scale and position, the existence of occlusion in real world scenarios, and a relative lack of suitably large, publicly available datasets for training.

For FER, CNN pipelines use full face images as inputs to the recognition pipeline. The aim of CNN learning is to derive discriminative feature sets which capture the entire range of features across the face. One way to achieve this is to utilise deep architectures which generate features with increasing levels of complexity. However, such architectures require large training sets and are computationally expensive and time consuming.

We can use behavioural performance from human visual perception tasks to enhance automated FER. There is evidence to suggest that humans find it easier to recognise happiness compared with other emotions (Calder, Keane, Young, & Dean, 2000; Calvo & Lundqvist, 2008; Calvo & Nummenmaa, 2009; Delicato, Finn, Morris, & Smith, 2014; Matsumoto & Hwang, 2014; Palermo & Coltheart, 2004; Recio, Schacht, & Sommer, 2013), and that they use a narrow band of low spatial frequencies for tasks such as face recognition (Keil *et al.*, 2008). Similar patterns regarding spatial frequencies have been identified when using machine learning for facial recognition (Keil *et al.*, 2008), and happiness is often more easily recognisable when applying deep learning for FER (Yu, 2015; Dachapally, 2017; Lopes, De Aguiar and Oliveira-Santos, 2015).

It has also been suggested that certain facial regions are more pertinent indicators of specific emotion categories (Bassili, 1979; Ellison & Massaro, 1997; Martin et al., 2012). Whilst some individuals consistently draw upon full face signals when recognising emotions, others can achieve comparable accuracy when viewing only the lower facial hemisphere (Delicato & Mason, 2015). We propose that by exploring these patterns and incorporating knowledge of how humans process emotion signals into CNN frameworks, one may achieve good FER accuracy whilst using relatively shallow networks (low spatial frequencies) and regional hierarchies. We evaluate the impact of integrating such knowledge into deep learning frameworks, report on classification accuracy, and discuss the implications of our findings.

This paper presents an evaluation of CNN learning for 6 class FER (Figure 1) using static images. The impact of architecture structure, preprocessing, and image input type is evaluated within a standardised evaluation protocol. Our contribution can be summarized as follows:

- Using findings from previously reported human visual perception tasks which explore FER across different regions of the face (Delicato & Mason, 2015), we propose a “human centric” CNN (HC-CNN) architecture for solving the FER problem.
- We quantitatively evaluate the performance of proposed HC-CNNs across a variety of architectures and parameterisations.
- We compare HC-CNN classification accuracy with existing behavioural patterns, thereby undertaking a qualitative analysis of human versus machine facial expression recognition.
- Finally, we combine multiple HC-CNN architectures into a classification hierarchy via an SVM meta layer, and demonstrate maximum classification accuracy of 93.3%. This constitutes an enhancement of 4.6% when compared against any of our single CNN models.

The remainder of this paper is organised as follows. A summary of related behavioural studies and related machine learning literature is provided in Section 2. Our research methodology is offered in Section 3. Results and conclusions are provided in Sections 4 and 5, respectively.

2. BACKGROUND AND MOTIVATION

2.1 Human Facial Expression Recognition

We showed previously that (Delicato & Mason, 2015) the features of a face that convey expressions vary depending upon the expression (happy or fear). When participants are asked to discriminate between a pair of faces, one of which was neutral and the other contained expression, they relied on the mouth to perform the task for happy expressions, while for fear they relied on the eyes. In some cases, participants were equally good at the task when the full face conveyed the happy expression, compared with when only the lower half of the image conveyed expression. This suggests that participants pay attention to different *regions* of the face depending upon the facial expression and supports the importance of featural processing in emotion recognition (see also Bombari, Schmid, Schmid Mast, Birri, Mast & Lobmaier, 2013). Based on these findings, we consider how one may use such behavioural data to inform the development of CNN architectures to classify emotions.

2.2 CNNs for Facial Expression Recognition

CNNs were first proposed by Lecun et al., (1998) and have been demonstrated as an appropriate

mechanism for learning high level feature abstractions for machine vision. Based on the premise that better recognition can be achieved via automatic feature learning as opposed to hand-crafted features, CNN architectures are formed as compositions of multiple layers, including convolutional, nonlinear, sub-sampling, and fully connected layers.

Convolutional layers may be regarded as feature identifiers, and are characterised by: the kernel size across which convolution is performed; and the number of generated feature maps. Nonlinear and subsampling layers provide nonlinearities and increase position invariance, to enhance model robustness and reduce overfitting. Subsampling layers are characterised by their kernel size (which defines the degree of dimensionality reduction performed) and step size (Lopes *et al.*, 2017). Common subsampling methods include maximum pooling and average pooling. The parameterisation and ordering of layer compositions can vary depending on the nature of the problem being explored. Given that outputs from shallow levels of a CNN act as inputs to deeper levels, the more convolutional layers added, the more complex the features detected and the richer the feature space used for classification.

For FER tasks, a variety of CNN architectures have been proposed and evaluated. The individual CNN architectures used for FER vary in terms of the number of layers, their parameterisation, and their computational cost. More recently, the application of ensemble CNN frameworks have also been investigated (Kim *et al.*, 2016).

Lopes, De Aguiar and Oliveira-Santos, (2015) apply a 7 layer CNN architecture with 2 convolutional layers (layer 1, 5*5 kernel; layer 2, 7*7 kernel), and 2 sub-sampling layers (2*2 filter, stride = 2). They evaluate the impact of normalization and synthetic data augmentation on classification accuracy using the extended CK dataset (CK+), and achieve average classification accuracy of 91.46% for 6 class CNN recognition (max 93.74%). Lopes, De Aguiar and Oliveira-Santos, (2015) found that augmenting training data through the addition of random perturbations and horizontal flips enhanced overall accuracy, a finding supported by Jung et al., (Jung, 2015). However, despite achieving high classification accuracy, the system proposed by Lopes, De Aguiar and Oliveira-Santos, (2015) requires the locations of each eye prior to image normalization. Shan *et al.*, (2017) evaluate a similar CNN architecture with two convolutional and two subsampling layers without eye location inputs and achieve a maximum classification of 76.4% using the JAFFE dataset and 80.63% using the CK + dataset.

Jung *et al.*, (2015) develop an image based deep recognition system with 3 convolutional layers (5*5 skernel), 3 subsampling layers (max pooling, 3*3 kernel, and stride = 2), and rectified linear unit (ReLu) activation. The procedure performs face detection and normalises image inputs to 64*64 pixels prior to convolution. Using the extended CK dataset for evaluation, where data is partitioned into 90% training and 10% testing sets, they achieve a maximum recognition rate of 86.54% across 7 classes of emotion, including neutral face. The least recognisable emotion was sadness, achieving true positive classification in only 44% of test cases.

Raghuvanshi and Choksi, (2016) train and evaluate 3 CNNs of varying depths and architectures for application to the Kaggle FER Challenge. The dropout rate, learning rate and regularisation values were all parameterised. Investigations on pooling approaches, use of batch normalisation and number of convolutional layers indicated that increasing dropout decreases overfitting a useful observation when working with smaller datasets. In cases of underfitting, increasing the number of fully connected layers can improve performance. Finally, Raghuvanshi and Choksi, (2016) observed that having a larger number of filters in deeper parts of the network led to higher accuracies. Overall, their best CNN achieved classification accuracy of 0.48 on the Kaggle blind test data.

Mousavi *et al.*, (2016) highlight the importance of developing CNN architectures which generalise across different datasets. They built a sparse 3-class system (positive, negative and neutral) with 3 convolutional layers (11*11 kernel, 10 filters per layer) and 3 maximum pooling layers (2*2 kernel, stride = 2) using the extended CK+ dataset. ReLu are applied to all convolutional layers. Mousavi *et al.*, (2016) evaluate performance using CK+ and when models generated using CK+ are applied to the JAFFE dataset (Lyons and Akamatsu, 1998). They achieve classification accuracy of 87% and 44% for CK+ and JAFFE, respectively.

2.3 CNN Ensembles for Facial Expression Recognition

In addition to the development of individual CNN models for FER, there exists a body of work focusing on the development of CNN ensembles and hierarchies (Liu, Zhang and Pan, 2016; Pramerdorfer and Kampel, 2016; Lopes *et al.*, 2017). Ensemble learning is an approach which combines multiple learned models with the objective of enabling better predictive performance, whilst reducing variance and overfitting. CNN ensembles have reported state of the art accuracy for CK+, and for challenges such as FER2013 and emotiW2015 (emotion recognition in the wild).

The winner of the EmotiW 2015 challenge utilized a large committee of CNNs (Kim *et al.*, 2016). To

ensure diversity, properties such as preprocessing, and convolutional kernel size were varied for each model. Predictions were integrated and network weights updated according to validation set performance. Liu, Zhang & Pang (2016) propose a network hierarchy which concatenates outputs from 3 CNNs of various architectural depths into a single set of features for softmax classification. Using zero mean centering and data augmentation, they achieve maximum classification accuracy of 65.03% on the FER2013 dataset. Kim *et al.*, (2016) implement an ensemble based method with varying networks and parameters. Using a hierarchical decision tree and an exponential rule to combine decisions of different networks they achieve accuracy of 61% in the EmotiW2015 task. Cu & Wong (2015) investigate classification accuracy when features extracted from VGG-16 and ResNet50 CNN architectures are combined and used within a logistic regression ensemble. When systems are evaluated on both the Kaggle and KDEF datasets, maximum (7 class) classification accuracy of 78.3% was achieved. Mollahosseini, Chan and Mahoor, (2015) evaluate CNN ensemble accuracy, and report state of the art performance, across a variety of datasets, including CK+. Their network consisted of two convolutional layers and 4 inception layers. For CK+, they report maximum classification accuracy of 93%.

The above findings suggest that ensembles and hierarchical architectures can outperform independent classifiers. This is also true when such approaches are combined with transfer learning. Yu, (2015) develops an architecture combining multiple deep CNNs via logistic regression. When pre-trained using FER 2013 challenge data and fine-tuned using Emotiw2015 data, the system achieved 61.29% accuracy in the 2015 static face expression in the wild (SFEW) challenge - a 22.16% improvement on the challenge baseline system.

3. METHODOLOGY

We implement the following classifiers from scratch:

1) A baseline network (CNN A) with one convolutional layer, as illustrated in Figure 2.
2) A CNN with 2 convolutional layers, a dropout layer, and 2 max pooling layers (CNN B), as illustrated in Figure 3.
3) A hierarchical architecture, which extracts features derived from multiple regional CNNs, and combines features via an SVM meta-layer.

Figure 1: *6 Class Prototypical Emotion Detection. From left to right: (top) happiness, sadness, surprise, (bottom) anger, disgust, fear. Source Extended CK Dataset* (Lucey *et al.*, 2010)

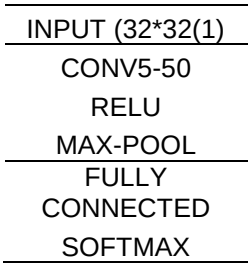

Figure 2: *Baseline CNN Architecture*

Figure 3: *CNN Architecture with 2 convolutional and 2 max-pool layers.*

3.1 Data

Similar to Lopes, De Aguiar and Oliveira-Santos, (2015), Shan *et al.*, (2017), and Mousavi *et al.*, (2016), we train and test our models using the extended Cohn- Kanade (CK+) dataset. CK + (Lucey *et al.*, 2010) is a popular dataset for performing automated facial expression recognition, which has been fully FACS coded and whose emotion labels have been validated. Image data is greyscale, and composed of 8 classes (happiness, fear, anger, contempt, disgust, surprise, sadness, and neutral) that contain sequences of facial expressions which start at a neutral face and end at the expression apex.

The subjects in the dataset are aged between 18 and 30, 35% male, 15% African American, and 3% Asian or South American. Similar to (Lopes, De Aguiar and Oliveira-Santos, 2015), we exclude contempt and neutral face from our evaluation and only include apex expressions for training and testing. Within this paper we therefore regard emotion recognition as a 6 class problem.

3.2 Preprocessing

Preprocessing steps include image resizing, zero mean centering, histogram equalisation, and data augmentation. Full-face image inputs are resized to 32*32 greyscale matrices, with the aim of reducing overfitting. To determine image input size we carried out preliminary investigations into the impact of input size on accuracy, using the baseline CNN. It was found that a reduction to 32*32 pixels resulted in no significant reduction in achievable classification accuracy.

3.2.1 Histogram Equalisation

Conventional image intensity values may be regarded as somewhat arbitrary. Images within a given dataset are not always comparable, even when they are of the same subject and when a standardized data acquisition protocol has been employed. Given that such variability will impact algorithm prediction and performance, it is desirable to normalize images to reduce variability in brightness and contrast.

Histogram equalisation is a simple but proven algorithm, which makes the greyscale distribution in a collection of images more uniform. As illustrated in Figure 4, the image input data can exhibit wide variations in intensity distribution, even for the same expression. After histogram equalisation against a reference image each histogram has a more uniform greyscale distribution and less variability.

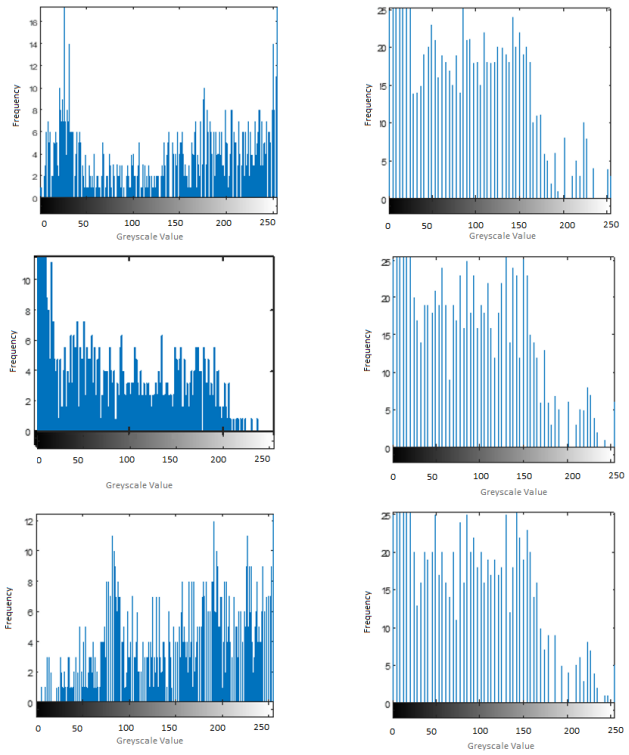

Figure 4: *Image histograms before (left) and after (right) histogram equalisation.*

3.2.2 Data Augmentation

Within the literature, it has been demonstrated that data augmentation constitutes an effective method for increasing deep learning accuracy when data volumes are low (Lopes *et al.*, 2017). Application of data augmentation increases the number of training samples, enhances the diversity of image inputs used during training, and can reduce the risk of overfitting (Liu, Zhang and Pan, 2016). We perform data augmentation via horizontal flip, as illustrated in Figure 5.

3.3 Architecture Parameterisation

For all individual CNN architectures, we use the following setup:

1) Initial learning rate = 0.0001.
2) Maximum number of epochs = 300.
3) Batch size = 10.
4) Convolutional kernel = 5*5, with 50 filters.

Mini-batch Gradient Descent is used to train networks, with batch size fixed to 10. Network weights are randomly initialized with numbers from a normal distribution. The learning rate is set to 0.0001. The maximum number of epochs is set to

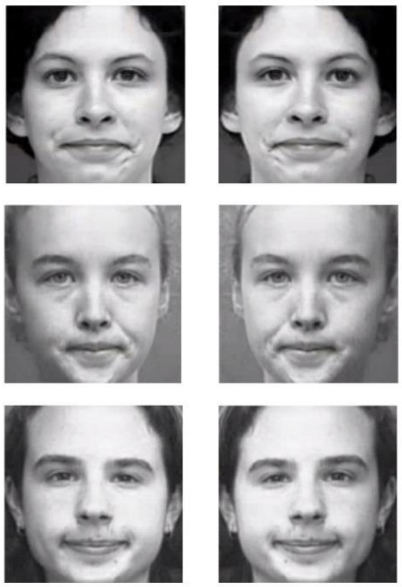

Figure 5: *Data Augmentation. Left: original image. Right: augmented image after horizontal flip.*

300, but training will terminate when the loss drops to a reasonable level.

3.4 Evaluation Protocol

We partition data into 90% training and 10% test sets, ensuring there is no participant overlap between groups. Training data is further partitioned in an 80-20 split, where 80% is used for CNN training and 20% for validation. Each CNN architecture is built and evaluated 100 times, with mean accuracy across all iterations recorded as a measure of performance. During data partitioning tasks, we balanced the data to ensure there were no differences of samples in parallel experiments. All experiments were carried out on a (Windows 7) virtual machine with 15 cores and 128GB RAM, with options to increase resources as required.

To explore the behavioural patterns described in Section 2, we compare performance when deep models are learnt both holistically and componentially. Specifically, we present accuracy achievable using:

1) The entire face region as image input (32*32 pixels).
2) The lower face hemisphere as classifier input (16 * 32 pixels).
3) The upper face hemisphere as classifier input (16 * 32 pixels).

Finally, to investigate whether learning facial hemispheres independently (and combining their outputs) can enhance classification accuracy, a multiple network fusion framework is presented. For each CNN architecture described above, we build two regional subnets, one for the lower facial hemisphere and one for the upper facial hemisphere. Resultant features are combined into a single classification module using an SVM meta-layer, with the choice of SVM upper layer being an initial default. We compare this framework against individual CNN results and against a CNN hierarchy which has two subnets, specifically the nets learned from full-face inputs for CNN A and CNN B respectively.

4. RESULTS

4.1 Comparison of Accuracies

Mean accuracy is evaluated (100 iterations) for all classifiers, using a) full face inputs, b) upper face inputs and c) lower face inputs. Given 6 emotion categories, random classification would give an accuracy of 16.67%. A description of accuracies for full face CNNs, lower face CNNs, upper face CNNS, and our classification hierarchy are presented individually below. In line with the majority of existing studies, we present mean accuracy as our principal performance metric. For each category of experiments (full face, upper, and lower analysis), we also present precision and recall metrics for the top performing classifier.

4.1.1 Full Face Analysis

Results of full face analysis for CNN architectures A and B are presented in Tables 1 and 2, where it is shown that maximum classification accuracy across all architectures was 88.7% for CNN B, trained with data augmentation. For both architectures, data augmentation enhanced performance, with classification accuracies of 85.5% and 88.7% for CNN A and CNN B respectively. Overall performance was better using CNN B for all experiments using preprocessing. CNN A with no preprocessing achieved the lowest accuracy of all full face models. The classification matrix for CNN B with data augmentation is illustrated in Table 3, where it is demonstrated that both happiness and fear were correctly recognised in 100% of test cases. The least correctly identified emotion was sadness, for which true positive classification was achieved in only 26% of test cases.

4.1.2 Lower Hemisphere Analysis

Results of lower face analysis for CNN architectures A and B are presented in Tables 4 and 5, respectively. It can be seen from Tables 4 and 5 that maximum lower hemisphere classification accuracy across all architectures was 87.37% for CNN A, with histogram equalisation. When analysing lower hemispheres, data augmentation did not increase

Table 1: *CNN A Results, Full Face Image Inputs.*

Pre-processing	Accuracy	
	Average (%)	*Maximum (%)*
None	*75.5*	*93.3*
ZM	*82.3*	*96.7*
HE	*82.8*	*93.3*
DA	*85.5*	*93.3*
All	*85.5*	*96.6*

Key: ZM, Zero Mean Centering; HE, Histogram Equalisation; DA, Data Augmentation; All, ZM + HE + DA.

Table 2: *CNN B Results, Full Face Image Inputs.*

Pre-processing	Accuracy	
	Average (%)	*Maximum (%)*
None	*82.3*	*90*
ZM	*88.1*	*96.7*
HE	*86.1*	*96.7*
DA	*88.7*	*96.7*
All	*88.7*	*96.7*

Key: ZM, Zero Mean Centering; HE, Histogram Equalisation; DA, Data Augmentation; All, ZM + HE + DA.

Table 3: *CNN B (DA) Classification Matrix, Full Face.*

	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	**0.77**	0.03	0.00	0.10	0.10	0.00
Disgust	0.03	**0.97**	0.00	0.00	0.00	0.00
Fear	0.02	0.00	**0.98**	0.00	0.00	0.00
Happy	0.00	0.00	0.00	**1.00**	0.00	0.00
Sadness	0.68	0.00	0.06	0.00	**0.26**	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	**1.00**

classification accuracy for either CNN architecture. The classification matrix for CNN A with histogram equalisation is shown in Table 6. It can be seen from Table 6 that surprise was correctly classified in 100% of cases, with happiness closely following and correctly identified in 99% of cases. The least frequently classified emotion was sadness, with true positive accuracy of 53%, and a misclassification of anger in 35% of cases.

4.1.3 Upper Hemisphere Analysis

Results of upper hemisphere analysis for CNN architectures A and B are presented in Tables 7 and 8. It can be seen from Tables 7 and 8 that models

Table 4: *CNN A Results, Lower Face Image Inputs.*

Pre-processing	Accuracy	
	Average (%)	*Maximum (%)*
None	*75.5*	*90*
ZM	*86.7*	*96.67*
HE	*87.37*	*96.67*
DA	*83.3*	*90*
All	*83.6*	*93.3*

Key: ZM, Zero Mean Centering; HE, Histogram Equalisation; DA, Data Augmentation; All, ZM + HE + DA.

Table 5: *CNN B Results, Lower Face Image Inputs.*

Pre-processing	Accuracy	
	Average (%)	*Maximum (%)*
None	*82.3*	*90*
ZM	*87.13*	*96.67*
HE	*86.97*	*96.67*
DA	*84.27*	*90*
All	*84.47*	*90*

Key: ZM, Zero Mean Centering; HE, Histogram Equalisation; DA, Data Augmentation; All, ZM + HE + DA.

Table 6: *CNN A (HE) Classification Matrix, Lower Face*

	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	**0.72**	0.00	0.00	0.00	0.18	0.10
Disgust	0.15	**0.81**	0.00	0.00	0.01	0.03
Fear	0.01	0.00	**0.98**	0.01	0.00	0.00
Happy	0.00	0.00	0.00	**0.99**	0.00	0.01
Sadness	0.35	0.00	0.07	0.00	**0.53**	0.05
Surprise	0.00	0.00	0.00	0.00	0.00	**1.00**

derived using upper hemisphere images achieve significantly lower classification rates than full face or lower hemisphere models, but still perform better than chance. Maximum classification accuracy across all architectures was 55.93% for CNN B, trained with data augmentation.

The classification matrix for CNN B with data augmentation is illustrated in Table 9. It can be seen from Table 9 that surprise and disgust were the most successfully classified emotions. The least accurate classification, which was only correctly identified in 1% of test cases, was fear. Using upper hemisphere inputs only, fear was incorrectly classified as anger in 50% of samples and misclassified as surprise in 49% of cases.

Table 7: *CNN A Results, Upper Face Image Inputs.*

Pre-processing	Accuracy	
	Average (%)	*Maximum (%)*
None	*46.5*	*66.67*
ZM	*52.5*	*66.67*
HE	*51.4*	*70*
DA	*51.4*	*73.3*
All	*54.03*	*70*

Key: ZM, Zero Mean Centering; HE, Histogram Equalisation; DA, Data Augmentation; All, ZM + HE + DA.

Table 8: *CNN B Results, Upper Face Image Inputs.*

Pre-processing	Accuracy	
	Average (%)	*Maximum (%)*
None	*45.5*	*63.3*
ZM	*48.9*	*63.3*
HE	*49.57*	*60*
DA	*55.93*	*66.67*
All	*55.77*	*63.33*

Key: ZM, Zero Mean Centering; HE, Histogram Equalisation; DA, Data Augmentation; All, ZM + HE + DA.

Table 9: *CNN B (DA) Classification Matrix, Upper Face*

	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	**0.29**	0.67	0.00	0.04	0.00	0.00
Disgust	0.18	**0.76**	0.00	0.06	0.00	0.00
Fear	0.50	0.00	**0.01**	0.00	0.00	0.49
Happy	0.00	0.31	0.00	**0.54**	0.00	0.15
Sadness	0.00	0.29	0.00	0.36	**0.34**	0.01
Surprise	0.00	0.01	0.00	0.21	0.00	**0.78**

4.1.4 CNN Hierarchy

Integrating upper and lower hemisphere models trained using CNN A and CNN B achieves classification accuracy of 93.3%, with 100% true positive classification of anger, fear, happiness, and surprise (Table 10). Building an equivalent hierarchy combining CNN A and CNN B outputs, but using only full face image inputs (Table 11), achieves classification accuracy of 90%. We have increased classification accuracy by 3.3 % simply by forcing (shallow) CNNs to independently learn facial hemispheres. Increase in accuracy relates to enhanced true positive classification of anger and disgust.

When comparing hierarchical approaches against single CNNs, evaluation of precision and recall further demonstrates improved performance (Table 12). Compared to the best individual CNN (full face, CNN B, with data augmentation), our upper and

Table 10: *Classification Matrix for hemisphere-model ensemble.*

	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	**1.00**	0.00	0.00	0.00	0.00	0.00
Disgust	0.17	**0.83**	0.00	0.00	0.00	0.00
Fear	0.00	0.00	**1.00**	0.00	0.00	0.00
Happy	0.00	0.00	0.00	**1.00**	0.00	0.00
Sadness	0.33	0.00	0.00	0.00	**0.67**	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	**1.00**

Table 11: *Classification Matrix for full face CNN ensemble.*

	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	**0.75**	0.00	0.00	0.00	0.25	0.00
Disgust	0.17	**0.83**	0.00	0.00	0.00	0.00
Fear	0.00	0.00	**1.00**	0.00	0.00	0.00
Happy	0.00	0.00	0.00	**1.00**	0.00	0.00
Sadness	0.33	0.00	0.00	0.00	**0.67**	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	**1.00**

Table 12: *Mean precision and recall (100 iterations) for top performing classifiers.*

Type	Model	Precision (%)	Recall (%)
Full Face	**CNNB (DA)**	84.30	83.30
Lower Face	**CNNA (HE)**	84.80	83.80
Upper Face	**CNN B (DA)**	61.00	45.30
Full Face	**Hierarchy**	88.80	87.50
Hemisphere	**Hierarchy**	94.40	91.70

lower hierarchy increases precision by 10.1% (from 84.3% to 94.4%) and recall by 8.4% percent (from 83.3% to 91.7%).

We have achieved high accuracy using relatively simple features (Figure 6). However, despite the high accuracy demonstrated using HC-CNN hierarchies, findings from individual subnets indicate that the positive benefits of data augmentation, cited elsewhere in the literature, have not been fully realised (Tables 4 and 5). This could be because a more aggressive augmentation approach, combining multiple random permutations of rotation per image, is required. Additionally, despite adding data augmentation, and dropout in the case of CNN B, we have not eliminated overfitting of data. This

becomes evident when training accuracies are compared with validation and test set accuracies (Figure 7). Overfitting is well known when working with small datasets, and future work requires exploration of potential solutions. A more thorough evaluation using a larger data set and novel augmentation approaches (such as synthesised faces) is necessary.

The relative poor performance of upper hemisphere models also compels additional evaluation. Potential reasons for observing low accuracies for upper hemisphere models are the extent of image subsampling applied during pre-processing, and the relatively shallow CNN architectures used for learning. We have previously shown that as facial expressions evolve temporally, they exhibit lower magnitudes of motion in the upper hemisphere region than in the lower facial hemisphere (Clawson et al., 2017). Within this research, image inputs are subsampled 32*32 matrices. To capture finer-grained upper face features, it may be necessary to utilise higher resolution image inputs and deeper CNN architectures.

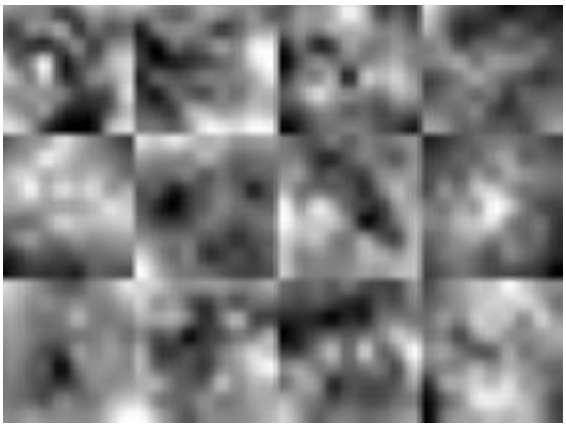

Figure 6: *Visualisations of CNN B features, Convolutional Layer 2.*

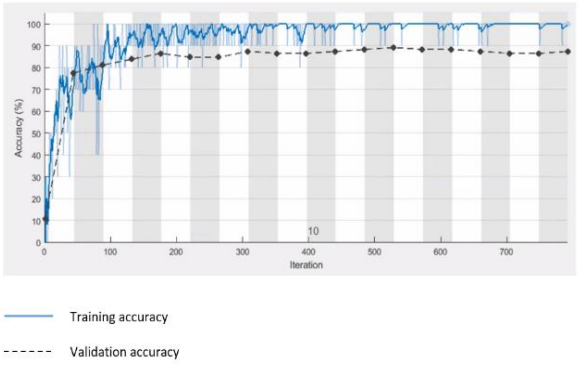

Figure 7: *Example training cycle, CNN B with Data Augmentation.*

4.2 Comparison with Behavioural Results

In a behavioural study we presented healthy participants with a pair of faces on a computer one after the other; one of the faces was always neutral and the other was expressive (Delicato, Wincenciak & Burn, 2016). The intensity of expression in the expressive face varied from neutral (0%) to full expression (100%). Participants were asked to indicate which of the two faces was "*more expressive*" the first or second using a computer input device. There were 6 different expressions presented; happy, angry, disgust, fear, sad and surprise.

As the intensity of the expression increased, participants found it easier to distinguish between the two faces and performance gradually increased from chance (for images with low intensity) to near 100% correct (for images with high intensity). We found that, on average, participants were more sensitive to expressions conveying happiness and least sensitive to expressions conveying sadness. The order of sensitivity (threshold intensity) to each universal emotion was happy (6%), surprise (8%), disgust (9%), fear (15%), anger (16%), sad (19%), where threshold is defined as the amount of intensity required to accurately distinguish between neutral and expressive faces on 75% of trials. If threshold is high, more expression in the face is required to perform the task well and this is described as low sensitivity. If threshold is low, less expression in the face is required to perform the task and this is described as high sensitivity.

This behavioural data is comparable to the data presented in this paper. The CNN B classification matrix (Table 3) shows that accuracy was greatest for Happy (1), Surprise (1), Fear (0.98) and Disgust (0.97) followed by Anger (0.77) and then finally Sad (0.26). With the exception of the response to Fear faces, the pattern of the output from the CNN is comparable to the behavioural data. The differences observed between performance of human participants and the CNNs may be related to processing of the affective content in the images rather than the visual information in the expression (see Calvo & Nummenmaa (2015) for a review).

In a second behavioural study, we looked at the sensitivity of individuals to faces where we manipulated the expression in the upper and / or lower part of the face for happy and fearful expressions (Delicato & Mason 2015). In addition to the full expressive face, five stimulus conditions were created. These stimuli were created by superimposing isolated expressive features (eye and / or mouth) from a full happy or fearful expression on the same actors neutral face (see Figure 13 expressive eyes (A), expressive mouth

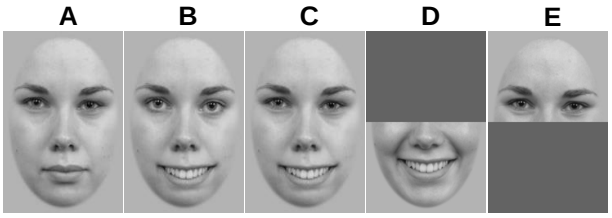

***Figure 13**: Example visual stimuli presented during behavioural study (Delicato & Mason, 2015).*

(B), expressive eyes and mouth (C), or by obscuring the upper or lower region of the face (eyes obscured (D), mouth obscured (E)).

We find increased sensitivity to happy expressions compared with fearful expressions for full face stimuli as well as expressive mouth (Figure 13.B), expressive eyes and mouth (Figure 13.C) and eyes obscured conditions (Figures 13.D). For expressive eyes (Figure 13.A) and mouth obscured (Figure 13.E) we find increased sensitivity to fearful expressions compared with happy expressions.

The output from CNNs presented in this paper would not predict such differences between happy and fearful expressions for both full face, lower or upper models. Indeed, for the CNN looking at upper hemispheres only, these network are better at classifying happy rather than fearful expressions.

5. CONCLUSIONS

In this paper we explored the task of facial expression recognition, and aimed to utilise behavioural knowledge of human visual perception to enable enhanced classification of images of faces. We proposed the use of regional inputs for CNN learning, experimented with single and hierarchical modelling approaches, and investigated the impact of image preprocessing and data augmentation. By forcing convolutional neural networks to learn specific facial regions independently, and combining their output with a SVM meta-layer, we achieved 93.3% accuracy on CK+ images.

Results are promising, and our methods could be widely applied across a range of HCI domains, including adaptive user interface development, usability testing, and mood tracking. Furthermore, we find similarities between overall trends in emotion detection when behavioural data is compared with CNN accuracy. Sadness is consistently more difficult to recognise in behavioural tasks and our research suggests that this may be because the emotion expressed is more subtle than other expressions. Work is currently ongoing to determine the relative contribution of the role of the strength of visual signal against individual sensitivity to affect.

Finally, we acknowledge the constraints of our research, and identify a need for further analysis. To this effect, future directions include: the extension of our system to consider eye and mouth regions of interest independently (as opposed to upper / lower hemisphere); the evaluation of alternative meta layer algorithms, including neural networks; and investigations into the development of CNN hierarchies incorporating data from multiple image resolutions.

We have used the CK+ dataset as an initial platform for analysis. There exists a need to evaluate our methods using multiple facial expression datasets, including more substantial 'in the wild' datasets. Furthermore, we aim to investigate the ability of our models to generalise to alternative datasets.

6. REFERENCES

Bassili, J. N. (1979). Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Journal of personality and social psychology, 37(11), 2049-2058.

Bombari, D., Schmid, P. C., Schmid Mast, M., Birri, S., Mast, F. W., & Lobmaier, J. S. (2013). Emotion recognition: The role of featural and configural face information. Quarterly Journal of Experimental Psychology, 66(12), pp. 2426-2442. doi:10.1080/17470218.2013.789065

Calder, A. J., Keane, J., Young, A. W., & Dean, M. (2000). Configural information in facial expression perception. Journal of Experimental Psychology: Human Perception and Performance, 26(2), 527-551.

Calvo, M. G., & Lundqvist, D. (2008). Facial expressions of emotion (KDEF): Identification under different display-duration conditions. Behavior Research Methods, 40(1), 109-115. doi:10.3758/BRM.40.1.109

Calvo, M. G., & Nummenmaa, L. (2009). Eye-movement assessment of the time course in facial expression recognition: Neurophysiological implications. Cognitive, Affective and Behavioral Neuroscience, 9(4), 398-411. doi:10.3758/CABN.9.4.398

Calvo, M. G., & Nummenmaa, L. (2016). Perceptual and affective mechanisms in facial expression recognition: An integrative review. [Article].

Cognition and Emotion, 30(6), pp. 1081-1106. doi:10.1080/02699931.2015.1049124

Cireşan, D., Meier, U. and Schmidhuber, J. (2012) 'Multi-column Deep Neural Networks for Image Classification', (February). doi: 10.1109/CVPR.2012.6248110.

Clawson, K., Delicato, L. S. & Garfield, S. (2017) Automated Representation of Non-Emotional Expressivity to Facilitate Understanding of Facial Mobility: Preliminary Findings. Intelligent Systems Conference, 7th – 8th September, London, 2017. doi: 10.1109/IntelliSys.2017.8324218

Dachapally, P. R. (2017) 'Facial Emotion Detection Using Convolutional Neural Networks and Representational Autoencoder Units'. Available at: https://arxiv.org/ftp/arxiv/papers/1706/1706.01509.pdf%0Ahttp://arxiv.org/abs/1706.01509.

Delicato, L. S., Finn, J., Morris, J. & Smith, B. (2014). Increased sensitivity to happy compared with fearful faces in a temporal two-interval forced-choice paradigm. European Conference on Visual Perception, Belgrade, Serbia. Perception 43(1), 75 doi: 10.1177/03010066140430S10

Delicato, L. S. & Mason, R. (2015) Happiness is in the mouth of the beholder and fear in the eyes. Vision Sciences Society, St. Petes Beach, Florida, US. Journal of Vision, Vol.15, 1378. doi:10.1167/15.12.1378

Delicato, L. S, Wincenciak, J., Burn, D. J. (2016). Evidence for a Face Inversion Effect in People with Parkinson's. Perception 45(S2), 295 doi: 10.1177/0301006616671273

Ellison, J.W. and Massaro, D.W., 1997. Featural evaluation, integration, and judgment of facial affect. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(1), p.213.

Jung, H. et al. (2015) 'Development of deep learning-based facial expression recognition system', 2015 Frontiers of Computer Vision, FCV 2015, pp. 2–5. doi: 10.1109/FCV.2015.7103729.

Keil, M. S. et al. (2008) 'Preferred spatial frequencies for human face processing are associated with optimal class discrimination in the machine', PLoS ONE, 3(7), pp. 1–5. doi: 10.1371/journal.pone.0002590.

Kim, B.-K. et al. (2016) 'Hierarchical committee of deep convolutional neural networks for robust facial expression recognition', Journal on Multimodal User Interfaces. Springer Berlin Heidelberg, 10(2), pp. 173–189. doi: 10.1007/s12193-015-0209-0.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., Gradient-based Learn. Appl. Doc. Recognit. 86 (11) (1998) 2278–2324, http://dx.doi.org/10.1109/5.726791.

Liu, K., Zhang, M. and Pan, Z. (2016) 'Facial Expression Recognition with CNN Ensemble', Proceedings - 2016 International Conference on Cyberworlds, CW 2016, pp. 163–166. doi: 10.1109/CW.2016.34.

Lopes, A. T. et al. (2017) 'Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order', Pattern Recognition. Elsevier, 61, pp. 610–628. doi: 10.1016/j.patcog.2016.07.026.

Lopes, A. T., De Aguiar, E. and Oliveira-Santos, T. (2015) 'A Facial Expression Recognition System Using Convolutional Networks', Brazilian Symposium of Computer Graphic and Image Processing, 2015–Octob, pp. 273–280. doi: 10.1109/SIBGRAPI.2015.14.

Lucey, P. et al. (2010) 'The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression', 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, (May), pp. 94–101. doi: 10.1109/CVPRW.2010.5543262.

Lyons, M. and Akamatsu, S. (1998) 'Coding Facial Expressions with GaborWavelets', third IEEE Conference on Automatic Face and Gesture Recognition, pp. 200–205. doi: 10.1109/AFGR.1998.670949.

Martin, D., Slessor, G., Allen, R., Phillips, L.H. and Darling, S., 2012. Processing orientation and emotion recognition. *Emotion*, *12*(1), p.39.

Matsumoto, D., & Hwang, H. C. (2014). Judgments of subtle facial expressions of emotion. Emotion, 14, 349–357. http://dx.doi.org/10 .1037/a0035237

Mollahosseini, A., Chan, D. and Mahoor, M. H. (2015) 'Going Deeper in Facial Expression Recognition using Deep Neural Networks'. doi: 10.1109/WACV.2016.7477450.

Mousavi, N. et al. (2016) 'Understanding how deep neural networks learn face expressions', Proceedings of the International Joint Conference on Neural Networks, 2016–Octob, pp. 227–234. doi: 10.1109/IJCNN.2016.7727203.

Palermo, R., & Coltheart, M. (2004). Photographs of facial expression: Accuracy, response times, and ratings of intensity. Behavior Research Methods,

Instruments & Computers, 36, 634–638. http://dx.doi.org/10.3758/BF03206544

Pramerdorfer, C. and Kampel, M. (2016) 'Facial Expression Recognition using Convolutional Neural Networks: State of the Art'. Available at: http://arxiv.org/abs/1612.02903.

Raghuvanshi, A. and Choksi, V. (2016) 'Facial Expression Recognition with Convolutional Neural Networks', pp. 1–8.

Recio, G., Schacht, A. and Sommer, W., 2014. Recognizing dynamic facial expressions of emotion: Specificity and intensity effects in event-related brain potentials. *Biological psychology*, *96*, pp.111-125.

Shan, K. et al. (2017) 'Automatic Facial Expression Recognition Based on a Deep Convolutional-Neural-Network Structure', IEEE Computer Society, pp. 123–128.

Sun, Y., Wang, X. and Tang, X. (2014) 'Deep Learning Face Representation from Predicting 10 , 000 Classes'.

Xie, S. and Hu, H. (2017) 'Facial expression recognition with FRR-CNN', Electronics Letters, 53(4), pp. 235–237. doi: 10.1049/el.2016.4328.

Yu, Z. (2015) 'Image based Static Facial Expression Recognition with Multiple Deep Network Learning', ACM on International Conference on Multimodal Interaction - ICMI, pp. 435–442. doi: 10.1145/2823327.2823341.